Decimation based algorithm to improve inference using the Pseudo-Likelihood

Aurélien Decelle[1] and Federico Ricci-Tersenghi [1,2]
[1] *Dipartimento di Fisica, Università La Sapienza, Piazzale Aldo Moro 5, I-00185 Roma, Italy.*
[2] *INFN-Sezione di Roma 1, and CNR-IPCF, UOS di Roma.*

In this Letter we propose a new method to infer the topology of the interaction network in a pairwise model with Ising variables. By using the pseudo-likelihood method (PLM) at high temperature, it is generally possible to distinguish between zero and non-zero couplings, because a clear gap separate the two groups. However at lower temperatures the PLM is much less effective and the result depends on subjective choices, as the value of the ℓ_1-regularizer and that of the threshold to separate non-zero couplings from null ones. We introduce a decimation procedure based on PLM, that recursively sets to zero the less significant couplings, until the variation of the pseudo-likelihood signals that relevant couplings are being removed. The new method is fully automatized and does not require any subjective choice by the user. Numerical tests shows that it performs better than standard PLM.

PACS numbers: 02.50.Tt,05.10.-a

INTRODUCTION

Recent years have seen a growing interest of the statistical physics community in the so-called 'inverse Ising problem' [1–7]. Although the problem is known since a long time ago under the name of Boltzmann machine learning [8], the renewed interest is linked to the large number of datasets coming from many different fields — e.g. biology, physics, neuroscience — that require a quantitative description. Building reliable models for these datasets has become therefore a fundamental problem [9–12]. Given a dataset and a model, the inverse problem aims to find the parameters of the model that best fit the data. Among all possible models, the Ising one, although very simple — it involves only pairwise interactions between discrete variables — can take into account a wide range of phenomena. It is thus natural to develop inference methods for this statistical model, as initially done by Boltzmann machine learning [8].

A common approach in Bayesian inference is to infer the model couplings by maximizing the likelihood function. Unfortunately, the likelihood depends on the partition function which is extremely difficult to compute in general. In this Letter we will therefore use the pseudo-likelihood approximation [13, 14] to perform inference on Ising models of reasonable size. This method has at least three advantages. First, it is possible to minimize the pseudo-likelihood function (PLF) in polynomial time for any Ising model. Second, the method is known to be exact in the case of infinite sampling [2]. Finally, as we will see in this Letter, the PLF can be used as an indication of how good the reconstruction is. It will allow us to decrease the number of model parameters progressively and to understand when it is worth stopping, thus leading to an accurate model selection algorithm.

We begin by introducing the pseudo-likelihood method (PLM) that has been already tested in inferring finite dimensional models and found to be very effective [2]. We also recall the ℓ_1-regularization extension and the thresholding procedure for discriminating non-zero couplings. We discuss the problem of very large coupling appearing in the PLM and how we solve this issue without using the ℓ_1-regularization. Finally, we describe the new decimation procedure to select accurately the model parameters and we show that the PLF can be used as a reliable likelihood. We conclude by showing that our new symmetrized PLM with decimation provides a very good inference of model couplings, better than standard PLM.

THE PSEUDO-LIKELIHOOD METHOD

The PLM is used in cases where the model parameters cannot be inferred correctly using mean-field methods and where the system size is too large to be able to maximize the true likelihood.

Let's consider an Ising model with probability measure

$$P_{\rm Boltz}(\underline{s}) = Z^{-1} \exp\left(\beta \sum_{i<j} J_{ij} s_i s_j + \beta \sum_i h_i s_i\right) , \quad (1)$$

whose parameters $\{J_{ij}, h_i\}$ we want to infer, given a set of M configurations $\{\underline{s}^{(k)}\}_{k=1\dots,M}$ independently sampled from $P_{\rm Boltz}(\underline{s})$. We define respectively the magnetizations and the correlations of the data as $m_i^{\rm D} = M^{-1}\sum_k s_i^{(k)}$ and $c_{ij}^{\rm D} = M^{-1}\sum_k s_i^{(k)} s_j^{(k)} - m_i^{\rm D} m_j^{\rm D}$.

Bayesian inference prescribes to maximize with respect to the model parameters the likelihood of the data

$$\mathcal{L} = \beta \sum_{i<j} J_{ij}(c_{ij}^{\rm D} + m_i^{\rm D} m_j^{\rm D}) + \beta \sum_i h_i m_i^{\rm D} - \log(Z) \ . \quad (2)$$

The maximum of $\mathcal{L}$ corresponds to parameters $\{J_{ij}^*, h_i^*\}$ matching the magnetizations and the correlations of the inferred model with those from the data

$$c_{ij}(\{J_{ij}^*, h_i^*\}) = c_{ij}^{\rm D} \ , \quad (3)$$

$$m_i(\{J_{ij}^*, h_i^*\}) = m_i^{\rm D} \ . \quad (4)$$

There are two general methods to find the parameters $\{J^*_{ij}, h^*_i\}$ that fulfill these conditions. The first one is by directly maximizing $\mathcal{L}$, but this requires to compute the partition function many times, which is impossible in practice. The second one, named Boltzmann machine learning [8], computes the left hand sides of eqs.(3,4) by Monte Carlo methods and updates the parameters as $J^{\rm new}_{ij} = \eta\big(c^{\rm D}_{ij} - c_{ij}(\{J^{\rm old}_{ij}, h^{\rm old}_i\})\big)$, where η is a learning factor. Again this requires to run long Monte Carlo simulations evaluating the average values with a precision good enough to sufficient to employ (3,4).

An important aspect in the inverse problem is therefore to find an approximation that provides a reasonably good estimate for the model parameters in a quick time (usually polynomial in the system size). A review of known approximations can be found in [15], to which we should add the more recent methods, as those based on the Bethe approximation [3, 4], the adaptive cluster expansion [1] and the probabilistic flow method [16]. In this Letter we compare our new decimation technique based on the PLM against the ℓ_1-regularization which is very much used to find the interaction topology [2, 14]. Standard mean-field method will not be considered as in general they perform weakly on finite dimensional systems, neither the adaptive cluster expansion nor the probabilistic flow method, since they strongly depends on several subjective choices (thresholds, dynamics, etc...).

Instead of using the likelihood function, which is very hard to compute, we use the PLF, $\mathcal{PL} = \sum_r \mathcal{L}_r$, where the index r runs over all variables, the "local" likelihood functions are given by

$$\mathcal{L}_r = \frac{1}{M}\sum_{k=1}^{M} \log p(s_r^{(k)}|\underline{s}^{(k)}_{\setminus r}) \,, \qquad (5)$$

and the conditional probability of variable s_r given the rest of the system is

$$p(s_r|\underline{s}_{\setminus r}) \equiv \Big[1 + \exp(-2\beta s_r(h_r + \sum_{j\neq r} J_{rj}s_j))\Big]^{-1} \,. \qquad (6)$$

The standard implementation of the PLM consists in maximizing each of the N local likelihood functions $\mathcal{L}_r$ separetely, thus getting two different estimates for each coupling J_{ij}: a first one J^{*i}_{ij} from the maximum of $\mathcal{L}_i$ and a second one J^{*j}_{ij} from the maximum of $\mathcal{L}_j$ (hereafter we assume $h_i = 0$ and $J_{ij} \in \{0, 1\}$ for the ease of presentation). Since the Ising model has symmetric couplings, the final estimate for the coupling J_{ij} is then obtained by taking the average, $J^*_{ij} = (J^{*i}_{ij} + J^{*j}_{ij})/2$.

This method has been confronted against mean-field methods for the SK model in [2] and it clearly estimates the couplings much better at low temperatures. For sparse models, its ability to infer correctly the interaction network (i.e. which couplings are non-zero) is largely improved by the use of the ℓ_1-regularization [14], thus maximizing the local functions $\mathcal{L}^{\ell_1}_r = \mathcal{L}_r + \lambda\sum_{j\neq r}|J_{rj}|$, with a suitably chosen (and not too large) λ regularizer. A further improvement in inferring the model topology has been achieved [2] by setting to zero all couplings whose estimate is below a threshold, $|J^*_{ij}| < \delta$ (but the choice of δ is delicate, as we discuss below).

Using the standard PLM we observe that, in difficult situations (e.g. when M is not large enough and temperature is low) some couplings are largely overestimated, $|J^*_{ij}| \gg 1$. In those cases, the inferred couplings are not symmetric ($J^{*i}_{ij} \neq J^{*j}_{ij}$) and only one of the two estimates is very large. The origin of this problem is to be found in the data information content, that is sometimes very poor around some variable s_i: in that case the estimates J^{*i}_{ij} are strongly unreliable (think e.g. to what happen if in the M samples s_i and its neighbors were almost always perfectly aligned). This problem is often partially solved by using the ℓ_1-regularization: the regularization parameter λ adds a weight on non-zero couplings and therefore prevents a coupling for being too large. A drawback of this approach is the tendency to underestimate globally the couplings and is therefore not completely satisfying. In this Letter, we choose a different solution: by maximizing the PLF $\mathcal{PL}$ (rather than the N functions $\mathcal{L}_r$ separately) we look for a compromise, where the estimate J^*_{ij} must be such that both $\mathcal{L}_i$ and $\mathcal{L}_j$ are reasonably large. The advantage of this maximization is that it provides a *unique* estimate for each coupling, which is in general of the right order of magnitude (unless the information content of the data is poor in a wide region). The PLF can be maximized by a standard Newton method, while for the ℓ_1-regularized functions we use the one of Ref. [17].

INFERRING TOPOLOGY USING PLM

In inference problems, the "model selection" is the ability to reduce the number of model parameters in order to keep only the essential ones (by Occam's razor rule the simplest model fitting a dataset is the one to be preferred). This model selection is both crucial to identify the structure underlying a dataset (e.g. the interaction network or the model topology) and to improve the quality of the coupling estimates (relevant couplings are better inferred after excluding the insignificant couplings from the model).

A common approach in the Ising inverse problem is to choose which are the null couplings by putting a cut-off on their estimated values. This is a reasonable choice as long as a clear gap separates the estimates of null couplings ($\boldsymbol{J}_0$) from those of non-zero couplings ($\boldsymbol{J}_1$), but it is not always the case. In the upper panel of Fig.1 we show an easy case, where a gap exists between estimates of $\boldsymbol{J}_0$ (red points) and those of $\boldsymbol{J}_1$ (blue points). However at lower temperature (upper data curve in the lower panel of Fig.1) the gap is completely absent and a good δ-thresholding to split $\boldsymbol{J}_0$ and $\boldsymbol{J}_1$ is impossible.

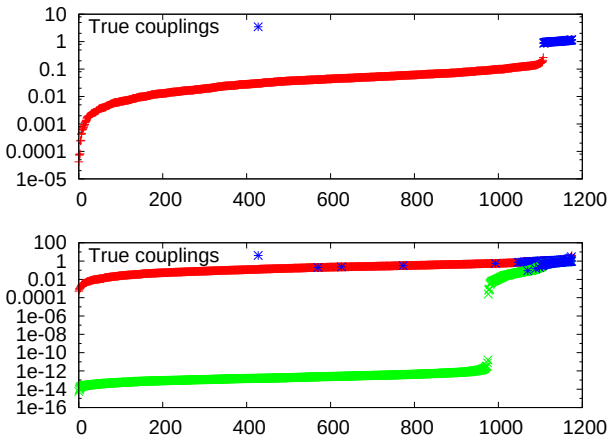

FIG. 1. (color online) One typical sample of a 2D ferromagnetic Ising model with 30% dilution ($M = 4500$). Top: PLM inferred couplings at $\beta = 0.5$ have a clear gap (blue crosses are $\boldsymbol{J}_1$ couplings). Bottom: PLM inferred couplings at $\beta = 0.9$ have no gap (upper data), while PLM+ℓ_1 shows a gap but does not recover perfectly the topology (lower data, $\lambda = 0.01$).

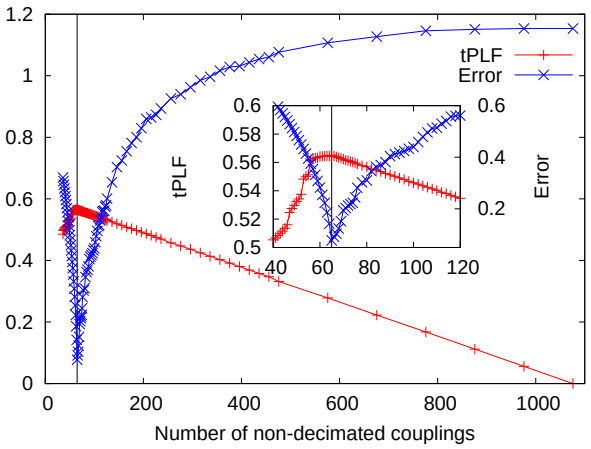

FIG. 2. Same model as Fig. 1 with $\beta = 0.9$ and $M = 4500$. The tPLF increases as the number of non-decimated couplings is reduced until it exhibits a maximum. In this case the maximum corresponds to a complete recovery of the graph topology and therefore to a small reconstruction error. The inset is a zoom on the maximum region.

For sparse models (as the one in Fig. 1) the use of the ℓ_1-regularization may induce a new gap (see lower data curve in the lower panel of Fig.1). However this ℓ_1-induced gap does not always separates correctly $\boldsymbol{J}_1$ from $\boldsymbol{J}_0$ couplings: many green points are above the gap in the lower panel of Fig.1. Moreover the estimates of $\boldsymbol{J}_1$ are systematically smaller than the true values, and this is unavoidable when using the ℓ_1-regularization. Finally, there is no consensus on how to choose the λ value.

We propose a new method for inferring non-zero couplings $\boldsymbol{J}_1$ that solves all the above problems. Our idea is to recursively set to zero couplings which are estimated very small by the PLM: we always maximize $\mathcal{PL}$ so as to avoid too large couplings and the bias due to the ℓ_1-regularization. At each decimation step we set to zero a finite fraction ρ of the remaining couplings, as such the total number of steps is $O(\log N)$ and the PLM+decimation algorithm is competitively fast. This fraction ρ is actually the only choice left to the user and the results are largely independent on it (in our tests we have used $\rho \leq 0.05$). Setting couplings to zero gradually is equivalent to use an adaptive thresholding with a very small δ value: so our new method should perform better than any standard thresholding procedure, especially because it does not require the existence of a gap in the inferred couplings (thus avoiding the use of ℓ_1-regularization that produces biased estimates and a strong λ-dependence).

The stopping criterion for the decimation procedure is based on the behaviour of the PLF. Indeed, we expect that, as long as the decimation procedure sets to zero couplings in $\boldsymbol{J}_0$ which are unnecessary to fit the data, the PLF should not change significantly. On the contrary, the pruning of a coupling in $\boldsymbol{J}_1$ should produce a drastic decrease in the PLF value. This expected behavior is confirmed by the numerical simulations. In practice we would like to stop the decimation at the point where the PLF variation, $\Delta\mathcal{PL}/\Delta n$ with Δn being the number decimated couplings in the last step, goes from 'small' to 'large' values. To make these two adjective quantitative we can compute the overall mean PLF variation during the decimation, that is the change in PLF between the fully connected model [where $\mathcal{PL}$ is maximized over all the $N(N-1)/2$ possible couplings and takes value $\mathcal{PL}_{\max}$] and the model of independent variables [no couplings left by the decimation and PLF equal to $-N\log(2)$]. The mean PLF variation is thus equal to $(\mathcal{PL}_{\max} + N\log(2))2/(N(N-1))$ and we propose to stop the decimation where the PLF variation reaches this mean value, that should separate 'small' from 'large' PLF variations. In practice it is more convenient to define the stopping point as the maximum of the *tilted PLF* (tPLF)

$$\mathcal{PL}^{\text{tilted}} \equiv \mathcal{PL} - x\mathcal{PL}_{\max} + (1-x)N\log(2)\ , \quad (7)$$

where x is the fraction of non-decimated coupling. It is easy to check that $\mathcal{PL}^{\text{tilted}} = 0$ both before starting the decimation ($x = 1$) and on a model with no coupling ($x = 0$). In the interval $[0, 1]$ a maximum appears if correlations are present in the dataset.

In Fig. 2 we show, for a case where inference is difficult, the tPLF as a function of the number of non-decimated couplings and the corresponding error in inferring couplings, defined as

$$\epsilon = \sqrt{\frac{\sum_{i<j}\left(J_{ij} - J^*_{ij}\right)^2}{\sum_{i<j} J^2_{ij}}} \quad (8)$$

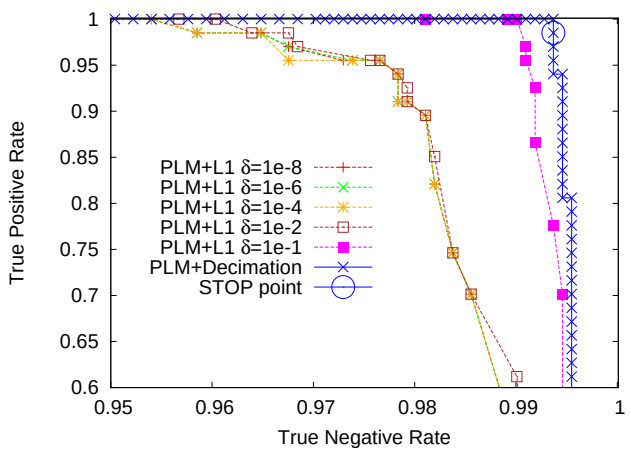

FIG. 3. The ROC graph for a typical sample of the 2D ferromagnetic Ising model with 30% of dilution at $\beta = 1.0$ and $M = 4500$ (difficult case). The upper right corner corresponds to the perfect reconstruction. For PLM+ℓ_1 curves are drawn by varying λ and they almost coincide if the threshold δ is correctly chosen in the gap, $\delta \in [10^{-8}, 10^{-2}]$. The better value $\delta = 0.1$ can not be chosen without previous knowledge about the topology. The PLM+decimation is clearly inferring topology better, even if PLM+ℓ_1 is finely tuned over λ and δ.

The zoom in the inset clearly shows that the maximum in the tPLF does corresponds to the minimum in the inference error. Please notice also as the density of data points along the curves changes, because we have decreased the value of ρ during the decimation in order to spend less time in the initial part (which is easy) and have denser points close to maximum (and thus improve its location).

Our new inference algorithm (PLM+decimation) can be very well compared to the standard PLM with ℓ_1-regularization and δ-thresholding by plotting the corresponding ROC curves (see Fig. 3). Each ROC curve is obtained by plotting parametrically the true positive rate (i.e. couplings in $\boldsymbol{J}_1$ inferred as non-zero divided by the total number of couplings in $\boldsymbol{J}_1$) versus the true negative rate (i.e. couplings in $\boldsymbol{J}_0$ inferred as null divided by the total number of couplings in $\boldsymbol{J}_0$). The ROC curve for an exact inference method run on noiseless data would pass through the upper right corner. In general an inference method is better the larger the area below the ROC curve. In the present case, the ROC curves for the standard PLM have been drawn by varying λ at fixed δ, while for the PLM+decimation the ROC curve has been drawn by varying the fraction of decimated couplings. Clearly the new method is outperforming standard PLM, even if the latter is optimized over λ and δ. To this respect, it is worth noticing that δ values in the range $[10^{-8}, 10^{-2}]$ do actually fall in the gap (see lower panel of Fig. 1) and lead to very similar ROC curves, while $\delta = 0.1$ is outside the gap and would be impossible to choose that value without knowing in advance the topology we are looking for. So the better ROC curve with $\delta = 0.1$ is practically unachievable and in general we believe that optimizing over δ should not produce any sensible improvement as long as δ is correctly chosen in the gap (to this respect the improvement obtained in Ref. [2] makes us suspect that a too large δ value was chosen based on the previous knowledge of the topology). So a fair comparison of our new method with the standard PLM should be made by choosing $\delta \leq 10^{-2}$ and the improvement is then very large. The stopping point selected by maximizing the tPLF is shown by a large dot in Fig. 3, and is indeed the closest to the upper right corner (full topology recover).

CONCLUSION

We have presented a new method for inferring the interaction topology of an Ising model which is based on the PLM and a decimation procedure that recursively sets to zero couplings which are inferred as the weakest. The procedure is fully automatized (apart from the choice of ρ which is mostly irrelevant for the results) and provides a unique answer to the inverse Ising problem. Execution times are comparable to those of standard PLM (i.e. polynomial in system size), apart from an extra $O(\log N)$ multiplicative factor (but remind that maximization without the ℓ_1-regularizer is easier). As the standard PLM, also our new method is exact in the large M limit, and shows for finite (and small) values of M much better performances than standard PLM with ℓ_1-regularization and δ-thresholding (which was considered among the best inference techniques available).

[1] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011).
[2] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).
[3] F. Ricci-Tersenghi, J. Stat. Mech., P08015 (2012).
[4] H. C. Nguyen and J. Berg, J. Stat. Mech., P03004 (2012); Phys. Rev. Lett. **109**, 050602 (2012).
[5] S.Cocco and R. Monasson, J. Stat. Phys. **147**, 252 (2012).
[6] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[7] J. Raymond and F. Ricci-Tersenghi, arXiv:1211.6400 (2012); arXiv:1301.1911 (2013).
[8] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Science **9**, 147 (1985).
[9] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[10] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).
[11] S. Cocco, S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. U.S.A. **106**, 14058 (2009).
[12] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc. Natl. Acad. Sci. U.S.A. **107**, 5405 (2010).
[13] J. Besag, The Statistician, **24**, 179 (1975).

[14] B. Ravikumar, M. Wainwright, and J. Lafferty, Ann. Stat. **38**, 1287 (2010).
[15] Y. Roudi, E. Aurell, and J. A. Hertz, Front. Comput. Neurosci. **3**, 22 (2009).
[16] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, Phys. Rev. Lett. **107**, 220601 (2011).
[17] K. Koh, S.-J. Kim, and S. Boyd, J. Mach. Learn. Res. **8**, 1519 (2007).